Static Stages for Heterogeneous Programming

Abstract

Heterogeneous architectures offer novel performance gains, but current programming models for orchestrating disparate hardware units fall short. CPU–GPU graphics applications, for example, use unsafe stringly-typed APIs for communication and brittle preprocessor macros for algorithmic specialization. We propose a new programming model for heterogeneous systems that improves safety and expressiveness without sacrificing performance. Programmers use *staging* annotations to express both code *placement* among the units in a heterogeneous system and code *specialization* for efficiency. Then, a *static staging* compiler generates code for each hardware unit and for the communication between them.

We implement static staging in BraidGL, a real-time graphics programming system. Programmers write hybrid CPU–GPU software in a unified language. The BraidGL compiler statically generates target-specific code and guarantees safe communication between the CPU and the graphics pipeline stages. Example scenes demonstrate the language's productivity advantages: BraidGL eliminates the safety and expressiveness pitfalls of OpenGL and makes common specialization techniques easy to apply.

1. Introduction

Heterogeneous computer systems are ubiquitous. Smartphone SoCs integrate dozens of disparate units; mainstream laptop and desktop chips include CPU and GPU cores on a single die; and even servers are beginning to embrace GPUs and FPGAs [34, 64]. Our work seeks to improve *heterogeneous programming,* where a single software artifact spans multiple computation units and memories with different resources, capabilities, ISAs, and performance characteristics.

Current techniques for heterogeneous programming tend toward two extremes. On one hand, low-level APIs such as CUDA [60] and OpenGL [70] offer precise control over when and where code executes. This close adherence to hardware interfaces yields performance, but it comes at the cost of safety and expressiveness. On the other hand, high-level abstractions that hide distinctions between hardware units, such as autotuning approaches [4, 49, 62] and domain-specific languages [16, 65], sacrifice low-level control over hardware components and communication between them.

Programmers should not need to sacrifice safety and expressiveness to gain control and performance. This paper proposes a new abstraction for heterogeneous programming that preserves direct control over hardware resources. We identify two fundamental concepts in heterogeneous programming: placement and specialization. *Placement* controls when and where code runs. Different units in heterogeneous systems have different capabilities and performance characteristics, so programmers need to control which code runs on which unit and how the components communicate. *Specialization* refines general code for efficient execution on specific hardware or inputs. Language support for specialization is crucial when optimizing code to fit the hardware's constraints. Graphics programmers, for instance, explore a torrent of algorithmic choices to extract high GPU performance while simultaneously producing a beautiful scene.

Staging for heterogeneity. We describe Braid, a language and compiler system that provides efficient abstractions for placement and specialization. In Braid, programmers map code to hardware units and describe how they communicate. The language defines *targets* that denote code for each architecture in the system. Braid's type system enforces safe communication and correct code specialization.

The key idea is to generalize work on *multi-stage programming* [74] to address both placement and specialization. Staging offers a foundation for safe interoperation between program components, but traditional staged languages focus on code generation and domain-specific optimization [12, 23, 66, 67]. Our compiler uses a new approach, *static staging*, to emit hardware-specific code and communication constructs ahead of time. Static staging lets Braid exploit classic staging's expressiveness without its run-time cost.

In Braid, stages represent both hardware places and specialization phases. With a *compile-time* stage, programmers rapidly explore strategies for specializing general code for the hardware target. Because both concepts rest on a consistent language foundation, specialization *composes* with placement in Braid: applications can use metaprogramming to decide which code to run and where. The unification and generalization of staging for high-performance heterogeneous programming is the key technical contribution in this work.

Braid's philosophy is a counterpoint to domain-specific languages that target exotic hardware [16, 65], where the goal is to hide low-level execution details from programmers. Instead, Braid exposes hardware details and makes them safe. Whereas DSLs are more appropriate for domain experts, Braid offers *system* experts a level of direct control over heterogeneous hardware that higher-level abstractions lack.

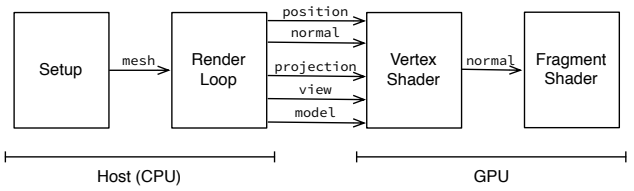


Figure 1. The stages in a simple graphics program. Each box is a stage and each arrow is a communicated value. The `position` and `normal` vector arrays describe a mesh. `projection`, `view`, and `model` are transformation matrices.

Case study in real-time graphics. Real-time graphics applications, such as video games, are arguably today's most widespread form of heterogeneous programming. They consist of software running on the CPU and in each stage of the GPU pipeline, but current APIs [56, 70] use separate languages in each context and unsafe interfaces between them. These applications also specialize code extensively for performance: modern games can produce hundreds of thousands of GPU program variants that customize visual effects for different scenes [35, 36]. To produce these variants, programmers must painfully explore alternatives with rudimentary token-stream manipulation, à la the C macro preprocessor.

We implement BraidGL, a real-time graphics programming system for hybrid CPU–GPUs and show how it addresses these problems with static staging. The compiler takes a single source program with staging annotations and emits a mixture of JavaScript for the CPU, GLSL for the GPU, and WebGL "glue code" [39]. Unlike traditional languages for shader programming such as OpenGL and Direct3D, BraidGL programs statically guarantee safe interaction: the communicating programs agree on the types of the values they exchange. Unlike general-purpose GPU languages that do not support shader pipelines such as CUDA and OpenACC [2, 60], BraidGL explicitly exposes the telescoping sequence of stages that make up the 3D rendering pipeline. Finally, unlike any other heterogeneous programming language we are aware of, BraidGL delivers safe compile-time and run-time metaprogramming tools that compose with staged execution.

We use three case studies to demonstrate that BraidGL combines performance, safety, and expressiveness. We show how to use static staging to explore specialization strategies that improve frame rates without adding complexity.

2. Overview by Example

This section illustrates the challenges of programming heterogeneous systems using WebGL [39], the JavaScript variant of OpenGL, a dominant API for GPU-accelerated 3D rendering. We then contrast current practices with Braid.

```
③ var VERTEX_SHADER =
    "attribute vec3 aPosition, aNormal;" +
    "uniform mat4 uProjection, uModel, uView;" +
    "varying vec3 vNormal;" +
    "void main() {" +
      "vNormal = aNormal;" +
      "gl_Position = uProjection * uView * uModel *" +
        "aPosition;" +
    "}",

④ var FRAGMENT_SHADER =
    "varying vec3 vNormal;" +
    "void main() {" +
      "gl_FragColor = abs(vNormal);" +
    "}"
  );

  // One-time setup.
① var program = compile(gl, VERTEX_SHADER, FRAGMENT_SHADER);
  var loc_aPosition =
    gl.getAttribLocation(program, "aPosition");
  /* ... */
  var mesh = loadModel();

  // Per-frame render loop.
  while (1) {
    // Bind the shader program and its parameters.
②   gl.useProgram(program);
    bindAttr(gl, loc_aPosition, getPositions(mesh));
    /* ... */
    draw(gl, getSize(mesh));
  }
```

Figure 2. Bare-bones WebGL to draw a mesh with a shader.

2.1. An Example in WebGL

In graphics programming, kernels called *shader programs* execute in a pipeline on the GPU to determine the appearance of objects in a scene. The two most common kinds are *vertex shaders*, which compute the position of each vertex in a 3D mesh, and *fragment shaders*, which compute the color of each pixel on an object's surface. The CPU sends parameter values to the vertex shader, which in turn may pass values to the fragment shader. Figure 1 illustrates these execution phases for the minimal example program in Figure 2 using JavaScript, WebGL, and GLSL. In this example:

- The *setup* phase ① loads 3D mesh data, compiles the GLSL shader source code, and installs "location" handles for CPU–GPU communication.
- The *render loop* ② executes on the CPU, once for each frame. It tells the driver which shader program to use, binds parameters using the location handles, and invokes a draw command. This example sends the standard transformation matrices `projection`, `view`, and `model` and the vector arrays `position` and `normal` from the mesh.
- The *vertex shader* ③ is a GLSL program that writes to the magic variable `gl_Position` to determine the coordinates of each vertex. This example multiplies the transformation matrices by the `position` vector from the mesh. It passes the `normal` vector to the fragment shader using the prefixed names `aNormal` and `vNormal` to distinguish the vertex and fragment shaders' views of the same data.

- The *fragment shader* ④ writes to `gl_FragColor` to produce each color's pixel.

This rendering code demonstrates the following pitfalls:

Programming in strings. The host program embeds the GPU programs as string literals and passes them to the graphics driver for just-in-time compilation. This *stringly-typed* API limits static checking and obscures data flow, causing syntax errors and undefined variables to manifest at runtime. The potential for subtle errors makes it awkward to experiment with code placement and optimization strategies.

Unsafe, verbose communication. Inter-stage communication involves three steps. Each shader declares its parameters and outputs: `uniform` denotes a per-draw-call parameter; `attribute` denotes a per-vertex parameter; and `varying` denotes a communication channel between different shaders. During setup, the host code uses each parameter's name to look up its handle. Finally, to draw each frame, the host code uses the handles to assign the parameter values. None of these constructs are type-checked at compile time: it is easy to mismatch types or use an undeclared variable.

Some general-purpose GPU programming languages, including CUDA and its successors [2, 38, 68], use a *single-source* model where communication can be checked at compile time. These simpler languages do not target real-time graphics programming, which requires composition of multiple nested stages to model the rendering pipeline.

Unscalable specialization. To extract high performance from the GPU, graphics applications specialize general code to create simpler shader *variants*. Because divergent control is expensive on GPUs, shaders avoid `if` branches. Instead, programmers often use GLSL's C-preprocessor-like `#ifdef` to generate cheaper, straight-line shader programs:

```
#ifdef TINT
gl_FragColor = abs(vNormal) + vec4(0.1, 0.1, 0.5, 1.0);
#else
gl_FragColor = abs(vNormal);
#endif
```

The C preprocessor approach suffices for a handful of build flags but does not scale well to the massive specialization required in modern games, which can generate hundreds of thousands of shader variants [35, 36]. Traditional `#ifdef` code manipulation makes it difficult to ensure that each variant will pass type checking or even parse. CUDA and other general-purpose GPU languages have the same limitation.

2.2. Real-time Graphics Programming with Braid

Braid uses static staging to address these problems. Figure 3 shows the BraidGL version of the WebGL program in Figure 2. The programmer *quotes* code with angle brackets `<`e`>` to control code generation and placement: `js<`e`>` delimits JavaScript code executing on the CPU and `glsl<`e`>` indicates GLSL code executing on the GPU. The example in Figure 3 uses three pairs of angle brackets to delimit four stages, from the outermost setup stage to the innermost fragment-shader

```
① var mesh = loadModel();

  # Render loop.
② render js<
    var position = getPositions(mesh);
    # ...

    # Vertex shader.
③   vertex glsl<
      gl_Position = projection * view * model * position;

      # Fragment shader.
④     fragment glsl<
        gl_FragColor = abs(normal);
      >;
    >;
    draw(getSize(mesh));
>;
```

Figure 3. The BraidGL equivalent of Figure 2's WebGL.

stage. Data flows from earlier to later stages using variable references that cross stage boundaries. The BraidGL compiler provides `render`, `vertex`, and `fragment` intrinsics to send code to the appropriate compute unit.

This section gives an overview of static staging via our example. Next, Section 3 describes Braid in detail.

Unified, type-safe programming. Braid does not use strings for code. The setup, render, vertex, and fragment stages all coexist in a single, type-safe program. This uniformity makes the flow of data clear, detects errors statically, and simplifies the job of moving computations between stages. For example, since the transformation matrices are constant in this program, we could move their multiplication to the CPU:

```
var pvm = projection * view * model;
vertex glsl<
  gl_Position = pvm * position;  # ...
```

This change reduces CPU–GPU communication but increases the CPU's workload, so the best placement depends on the platform. A uniform programming model simplifies rapid exploration of placement strategies.

Safe communication. Braid enforces a consistent interface between heterogeneous components. The vertex shader references the `position` variable, which is defined in the render-loop stage. Cross-stage references are a special case of a more general concept in Braid called *materialization* that abstracts communication channels in heterogeneous hardware.

Metaprogramming for specialization. Programmers can define *staging-based macros* that transform code for efficiency. A programmer might write a compile-time conditional macro, `@static_if`, to replace the `#ifdef`-based conditional tinting from above:

```
gl_FragColor =
  @static_if tint
    (abs(normal) + vec4(0.1, 0.1, 0.5, 1.0))
    abs(normal)
```

The key idea is to treat the compilation phase as another stage that precedes runtime setup. Stages thus offer a common basis for compile-time specialization and run-time coordina-

tion. Unlike `#ifdef` metaprogramming, Braid's safety guarantees ensure that specialized shader code is well-typed.

2.3. Dynamic vs. Static Stages

Braid's static staging approach generalizes *multi-stage programming* [74]. Traditional staging focuses on constructing compilers and domain-specific languages [12, 23, 66, 67]. In this *dynamic staging* approach, a stage-0 program generates a stage-1 program, which may generate a stage-2 program, and so on. If the stages need to communicate, an earlier stage bakes values into the later-stage program as constants [44, 66, 74]. The type system guarantees that all generated programs will type check. Implementations, such as Scala LMS [66], Terra [23], and MetaOCaml [14], build up abstract syntax trees and compile them on the fly. The AST-based approach incurs overheads for manipulating and generating code at run time, but the costs are acceptable because only the *compiler* uses the staging constructs, not the application. Our work delivers the power of staging to application developers.

The key difference in static staging is that the compiler generates code ahead of time for all stages. Programs use stages to mark the boundaries between systems, and the static approach ensures that this separation does not come with a run-time cost. To meet this goal, static staging introduces *materialization*, a language construct to denote inter-stage communication. Materialization coexists with traditional *splicing*, which retains prior work's code generation semantics.

3. Static Staging in Braid

Braid's goals are to express placement and specialization in a heterogeneous system while statically compiling efficient code. Its unique features in service of these goals are:

- A new kind of escape expression, called *materialization*, that expresses inter-stage communication.
- Multi-level escape expressions, which compose staging for specialization and staging for placement.
- An optimization, *pre-splicing*, that avoids the cost of run-time metaprogramming while preserving its semantics.
- A hygienic macro system that uses static staging to encapsulate reusable specialization strategies.

Table 1 enumerates Braid's staging constructs.

This section describes the Braid language and its prototype compiler. For simplicity, the basic Braid system has only one target, JavaScript, and uses unannotated quotes `<`e`>`. The compiler emits string-wrapped JavaScript code for each `<`e`>` and runs the code using `eval`. Section 5 shows how we add compiler backends for hardware targets in real-time graphics.

3.1. Traditional Staging: Quote, Run, and Splice

Braid borrows three central constructs from traditional multi-stage programming: the quote, run, and splice expressions. *Quotation* (also known as *bracketing* in MetaML [74] and *quasiquotation* in Lisp [8]) denotes a deferred computation and produces a *code value*. In Braid, quotation is written using angle brackets:

```
var program = <
  var highlight = fun color:Vec3 ->
    min(color * 1.5, vec3(1.0, 1.0, 1.0));
  highlight(vec3(0.2, 0.8, 0.1))
>
```

Here, `program` is a code value that, when executed, defines and invokes a function. Its type is written `<Vec3>`. A *run* operator, `!`e, executes code values. Here, `!program` evaluates to the `Vec3` value $(0.3, 1.0, 0.15)$.

Quotations use a *splice* expression (also called *escape* or *unquote*), to compose code values. In Braid, square brackets denote splicing. The splice expression `[`e`]` must appear inside a quote. It evaluates e in the quote's *parent* context to produce a code value to insert into quote. For example, splicing can "bake in" a parameter as a constant:

```
var makeHighlight = fun amount:<Float> ->
  < fun color:Vec3 ->
      min(color * [amount], vec3(1.0, 1.0, 1.0)) >
```

The `[amount]` splice looks up a code value from the environment *outside* the quote and splices into the body of the quoted highlighting function. Running the spliced quote with `!(makeHighlight(<2.0>))` produces a function with the literal `2.0` inlined into its body. The result is equivalent to:

```
fun color:Vec3 ->
  min(color * 2.0, vec3(1.0, 1.0, 1.0))
```

In this example, splicing is akin to partial evaluation.

Compiling stages statically. The baseline Braid compiler emits JavaScript code for all stages. It wraps quoted code in string literals and compiles the run operator `!` to JavaScript's `eval`. Splicing uses string replacement with a magic token. Our `makeHighlight` function compiles to this JavaScript:

```
var QUOTE_1 = "function (color) {" +
"  return min(color * __SPLICE_1__, vec3(1.0, 1.0, 1.0))" +
"}";
var makeHighlight = function (amount) {
  return QUOTE_1.replace("__SPLICE_1__", amount); }
```

Our above invocation compiles to JavaScript using `eval`:

```
var QUOTE_2 = "2.0";
var highlight2 = eval(makeHighlight(QUOTE_2));
```

In Braid, all stages are compiled to executable code; there is no runtime AST data structure. This ahead-of-time code generation is key to performance in heterogeneous programming. BraidGL, for example, emits complete GLSL programs from shader quotes at compile time so it can match the performance of manually-written OpenGL code.

This trio of staging concepts—quote, run, and splice—make Braid a safe and efficient tool for traditional metaprogramming. The rest of this section describes Braid's extensions to staging for heterogeneous programming.

3.2. Expressing Communication: Materialization

Braid adds an abstraction for efficient inter-stage communication. In traditional dynamic staging, the only way to communicate with a later stage is to *lift* a raw value to a code value

Syntax	Name	Section	Purpose
`target<e>`	quote	3.1	Delay computation of `e` and return a code value for `target`.
`!e`	run		Execute a code value.
`[e]`	splice		Inline a code value from the current quote's parent scope.
`%[e]`	materialize	3.2	Communicate a value from the parent scope.
`n[e]`	multi-level escape	3.3	Evaluate `e` in the context `n` levels away and splice.
`$[e]` & `$<e>`	open code	3.4	Splice while preserving the current quote's scope.
`@name`	macro invocation	3.5	Call a function defined at any earlier stage.

Table 1. Language constructs in Braid.

and splice it into the later stage's code [74]. Traditional splicing does not suffice for heterogeneous programming, where it is both inefficient and inflexible. Instead, Braid introduces a new communication construct, called *materialization*, that makes a value from one stage available in another.

For example, this quote denotes a GPU-side function that adjusts a color by a CPU-specified amount:

```
var amount = ...;
< fun color:Vec3 ->
    min(color * %[amount], vec3(1.0, 1.0, 1.0)) >
```

The expression `%[amount]` is a materialization escape. Like a splice escape, `%[`e`]` evaluates e in the current quote's parent context. But unlike splicing, materialization communicates e through shared memory or dedicated communication channels—it does not manipulate any code. In BraidGL, materialization abstracts the OpenGL APIs for binding shader parameters. Materialization distinguishes Braid from prior multi-stage languages, which view splicing as the only method for cross-stage persistence [33, 44].

Implementing materialization. The Braid implementation compiles quotes with materialization using a strategy similar to compiling closures with lambda lifting. A code value consists of a code pointer and an environment containing materialized values. Our highlighter example compiles to:

```
var QUOTE_1 = "... color * m1 ...";
{ code: QUOTE_1, env: { m1: amount } }
```

The code value is a JavaScript object. Its `env` member maps the opaque materialization name `m1` to a concrete value. To execute code values with materialized data, programs bind the environment's names and then invoke `eval`.

Compiler extensions define how to implement materialization when it traverses heterogeneous targets. Section 5 describes how BraidGL implements materialization for CPU–GPU communication.

Cross-stage references. In Braid, later-stage code may refer to variables defined at earlier stages. These references use materialization under the hood but eliminate some annotation burden. For example, this version of the highlighter program uses cross-stage references to `amount` and `color`:

```
fun color:Vec3 amount:Float ->
  < min(color * amount, vec3(1.0, 1.0, 1.0)) >
```

This code has identical semantics to a quote that uses explicit materialization escapes:

```
< min(%[color] * %[amount], vec3(1.0, 1.0, 1.0)) >
```

The type checker keeps track of the stage where each variable is defined, and when a reference crosses a stage boundary to reach its definition, the compiler uses materialization to communicate the value.

Cross-stage references are *nearly* syntactic sugar for materialization escapes, but they differ slightly when a stage refers to the same variable multiple times. Consider the highlighter with a parameterized channel limit:

```
< min(color * amount, vec3(limit, limit, limit)) >
```

This code should not waste bandwidth to communicate three copies of the `limit` value to the GPU. The Braid compiler therefore ensures that cross-stage references communicate their values only once.

3.3. Multi-Level Escapes

The two roles of staging in Braid, specialization and placement, must compose safely. Specifically, programs may specialize quotes that use staging for placement. For example, this program creates code to run on the GPU:

```
var amount = ...;
var shader = fun color:Vec3 -> <
  gl_FragColor = [
    if (amount == 1.0)
      <color>
      <min(color * amount, vec3(1.0, 1.0, 1.0))> ] >
```

It uses a splice escape to decide whether to emit code to highlight the color parameter. When `amount` is 1.0, the shader can use simpler code and avoid wasting bandwidth to communicate the value. However, this program uses *runtime* metaprogramming: every invocation of the `shader` function will splice together a fresh shader code string. This dynamic code generation is unacceptably inefficient in a graphics system that needs to draw thousands of objects per frame.

When `amount` is a static parameter, Braid provides tools to instead perform all the metaprogramming at compile time. An ordinary splice escape only enables code generation at the immediately preceding stage—in this case, the CPU execution stage. Braid generalizes this concept to skip through multiple stages and perform metaprogramming at any stage. By annotating an escape expression `n[`e`]` with a number n, programs can evaluate e and splice it at the nth prior stage *without passing through the intermediate stages*. For exam-

ple, we can introduce a compile-time stage by wrapping our program in a new, top-level quote:

```
var amount = ...;
!<
   var shader = fun color:Vec3 -> <
      gl_FragColor = 2[
         if (amount == 1.0)
            <color>
            <min(color * amount, vec3(1.0, 1.0, 1.0))> ] >
>
```

By writing the splice escape as `2[...]`, this version performs splicing at the *outer* specialization stage. Executing this program splices the final program, but this output code is complete. It does not contain any splicing itself. The shader code it contains uses one of the two alternative expressions and omits the other entirely.

Implementing generalized splicing. The Braid compiler recursively emits nested quotes using nested JavaScript string literals. Earlier-stage splicing works by replacing a token within an inner string:

```
var amount = ...;
var QUOTE_1 = "var shader = function (color) {" +
   "var QUOTE_2 =" +
   "  \"gl_FragColor = __SPLICE_1__\";" +
   "return QUOTE_2;" +
"}";
QUOTE_1.replace("__SPLICE_1__", ...);
```

The two-level splice `2[`e`]` is *not* equivalent to a nested splice `[[`e`]]`. The latter splices code twice, once in each stage:

```
var amount = ...;
var QUOTE_1 = "var shader = function (color) {" +
   "var QUOTE_2 =" +
   "  \"gl_FragColor = __SPLICE_1__\";" +
①  "return QUOTE_2.replace(\"__SPLICE_1__\", __SPLICE_2__);"
 + "}";
② QUOTE_1.replace("__SPLICE_2__", ...);
```

The outer splice ② runs first and inserts code into the outer quote, which splices again ① to transfer the value into the inner quote. Multi-level splicing is crucial to generating staged code that does not itself use runtime metaprogramming.

3.4. Expressive Metaprogramming with Open Code

As described so far, quotes evaluate to self-contained, executable programs. They cannot contain unbound variable references, which rules out important specialization strategies. For example, a graphics shader might conditionally darken or lighten a computed color:

```
var night = ...;
var shader = <
   var color = get_color();
   gl_FragColor = [
      if night
①        <color * 0.7>
②        <min(color * 1.2, vec3(1.0, 1.0, 1.0))> ] >
```

but the quoted `color` references ①② are illegal because the variable is not defined in any earlier stage.

In the multi-stage programming literature, self-contained quotes with no out-of-scope references are called *closed code*. Some languages allow *open code*, where quotes can refer to variables that are locally undefined [9, 57, 59, 73]. Unconstrained open code prohibits static code generation because variable references can be undefined until run time. Braid introduces a limited form of open-code quotation that allows static compilation.

In Braid, programs opt into open code by annotating escapes and quotes with a `$` prefix. For instance, this version of the quoted shader code is legal:

```
var color = get_color();
gl_FragColor = $[
   if (night)
      $<color * 0.7>
      $<min(color * 1.2, vec3(1.0, 1.0, 1.0))> ]
```

Open quotes `$<...>` share the environment of their nearest corresponding `$`-prefixed escape expression, so the references to `color` in this version of the code are legal.

To generate efficient code for open-code quotes, the compiler needs to know their surrounding context. The type system enforces this requirement by associating each open quote with its unique splice expression and making it an error to splice it anywhere else. This simple type constraint, where every quote has exactly one splice point, suffices, but future work may explore rules that add flexibility while still enforcing static compilation. For example, the compiler could allow multiple potential splice points, as long as they are statically enumerable, or programs could transform open quotes with additional context before splicing them at their destinations.

3.4.1. Pre-Splicing to Avoid Code Generation

Braid's restrictions on open code provide an opportunity to avoid the cost of runtime metaprogramming where it seems unavoidable. Consider a version of our highlighting shader where `amount` is a runtime parameter:

```
var shader = fun color:Vec3 amount:Float -> <
   gl_FragColor = $[
      if (amount == 1.0)
         $<color>
         $<min(color * amount, vec3(1.0, 1.0, 1.0))> ] >
```

Because `amount` is unknown at compile time, the multi-level escape approach from Section 3.3 will not work. Instead, this program splices GLSL code just before executing it to avoid a branch on the GPU. But there are only two possible final versions of the complete shader program, so it is wasteful to re-splice the code on every `shader` invocation.

The Braid compiler introduces an optimization to avoid wasteful code generation. *Pre-splicing* leverages the type system's open-code restrictions, which ensure that the set of resolutions for each `$[...]` escape is statically enumerable. The optimization iterates through each possible resolution of each such escape and produces a *variant* of the quote that inlines the chosen quotes. Then, it transforms the code from each splice escape to look up the correct variant in a table. In our example, pre-splicing produces optimized code equivalent to this escape-free program:

```
var shader = fun color:Vec3 amount:Float ->
   if (amount == 1.0)
      < gl_FragColor = color >
      < gl_FragColor = min(color * amount, vec3(1.0, 1.0, 1.0)) >
```

The JavaScript backend emits a switch on the IDs that identify each resolution:

```
var QUOTE_1_1 = "...";   // The pre-spliced variants.
var QUOTE_1_2 = "...";
var id = (amount === 1.0) ? 1 : 2;
switch (id) {   // Variant lookup.
case 1:
  return QUOTE_1_1;
case 2:
  return QUOTE_1_2; }
```

Pre-splicing trades reduced runtime cost for increased code size by generating a combinatorial space of specialized programs. For example, if a quote has two pre-spliced escapes, and each contains 3 possible subquotes, then the optimization will produce 3×3 complete variants.

3.5. Reusable Specialization with Macros

To make specialization strategies reusable, Braid adds a staging-based macro system. Library writers express common patterns for optimizing general code, and application writers reuse them without too much careful reasoning about stages. For example, some examples above use an `if` at an earlier stage to statically choose between two expressions. Using Braid's macro system, we define a "specialized `if`" construct, `@spif`, which substitutes for `if` in quoted code:

```
var night = ...;
< var value = @spif night 0.3 0.8;
  gl_FragColor = vec3(value, value, value) >
```

In Braid, macro invocations are syntactic sugar for explicit staging constructs. A macro invocation, written `@name`, invokes the function `name` at the stage where `name` is defined and splices the result into the current stage. There is no special syntax to define a macro; any function that accepts code values as arguments can be a macro. To define `@spif`, for example, a library author writes an ordinary function that takes code values as arguments:

```
var spif = fun cond:<Bool> t:<Float> f:<Float> ->
  if !cond t f;
```

The Braid compiler desugars macro invocations to multi-level splice escapes (Section 3.3) that skip to the stage where the function's name is bound. The macro's arguments become quote expressions. In our example, the macro is only one quotation level away, so the `@spif` invocation desugars to a single-level escape:

```
1[ spif <night> <0.3> <0.8> ]
```

The macro syntax offers a convenient way for Braid programmers to harness the power of multi-level splicing for specialization without directly reasoning about stages.

Macro functions can also take open-code quotes as arguments. Our `@spif` macro, for example, can let its `t` and `f` arguments refer to variables in the calling context, as in:

```
< var color = ...;
  var value = @spif night (0.3 * color) color; >
```

To resolve the references to `color`, the `spif` definition uses open-code arguments marked with `$`:

```
var spif = fun cond:<Bool> t:$<Float> f:$<Float> ->
  if !cond t f
```

$$
\begin{aligned}
e ::= {} & c \mid x \mid \textbf{var}\, x = e \mid e\,;\, e \\
& \langle e \rangle \mid\, !e \mid {}_n[e] \mid \%[e] \mid {}^{\$}_n[e] \mid \$\langle e \rangle \\
& x \in \text{variables},\ c \in \text{constants},\ n \in 1, 2, \ldots
\end{aligned}
$$

Figure 4. Syntax for BraidCore.

The `$<Float>` parameter type indicates that `t` and `f` should be passed as open quotes.

Braid's type system for staging ensures that code remains well-typed in all stages, and the syntactic sugar for macro invocations inherits the same guarantee. In other words, Braid's macro system is *hygienic:* macro writers need not worry about aliasing names in specialized code [47, 48].

4. Type System and Semantics

We formalize a semantics for BraidCore, a minimal version of Braid, to rigorously define its staging constructs and safety guarantees. This section summarizes the formalism; an accompanying technical report gives the full static and dynamic semantics (citation omitted for blind review). Figure 4 lists the abstract syntax for BraidCore.

The static staging semantics for BraidCore are simpler than most traditional staging semantics because of the limitations on open code. Approaches such as MetaML [74] use a tagging strategy: each value carries an integer tag that indicates its relative stage number (negative for earlier stages, positive for later stages). BraidCore instead organizes values into per-stage environments on a stack where variables always resolve in the topmost environment.

Type system. Types are either primitive or code types:

$$\tau ::= t \mid \langle \tau \rangle \qquad t ::= \texttt{Int} \mid \texttt{Float} \mid \cdots$$

A type context Γ consists of a stack of per-stage contexts γ that map variable names to types:

$$\Gamma ::= \cdot \mid \gamma, \Gamma \qquad \gamma ::= \cdot \mid x : \tau, \gamma$$

The typing judgment $\Gamma_1 \vdash s : \tau; \Gamma_2$ builds up a context. The rule for variable lookup retrieves a value from the γ on the top of the stack, and assignment adds a new mapping:

$$\textsc{type-lookup}\ \frac{}{\gamma, \Gamma \vdash x : \gamma(x); \gamma, \Gamma} \qquad \textsc{type-var}\ \frac{\Gamma_1 \vdash e : \tau; \gamma, \Gamma_2}{\Gamma_1 \vdash \textbf{var}\, x = e : \tau; (x : \tau, \gamma), \Gamma_2}$$

Cross-stage references are omitted from BraidCore because they can be modeled as syntactic sugar for materialization. A splice expression ${}_n[e]$ checks the expression n levels up:

$$\textsc{type-splice}\ \frac{\Gamma_1 \vdash e : \langle \tau \rangle; \Gamma_2 \qquad \text{len}(\overline{\gamma}) = n}{\overline{\gamma}, \Gamma_1 \vdash {}_n[e] : \tau; \overline{\gamma}, \Gamma_2}$$

Here, $\overline{\gamma}, \Gamma$ denotes a prefix $\overline{\gamma}$ and a tail context Γ, and $\text{len}(\overline{\gamma})$ determines the prefix's size. The result of e must be a code type. Materialization is similar, but the escape only moves a single stage and e need not result in a code type.

Dynamic semantics. We define a big-step operational semantics. Following the type system, a heap H consists of a stack of per-stage environments h. Values are either constants c or code values $\langle\!\langle e, h \rangle\!\rangle$, which consist of an expression and an associated environment that contains the results of materialization expressions.

The main big-step judgment $H; e \Downarrow H'; v$ evaluates an expression to a value and updates the heap. As with the type system, we use $h(x)$ to denote variable lookup. The rules for assignment and variable lookup work with the top per-stage environment in the heap:

$$\textsc{lookup}\quad \frac{}{h, H; x \Downarrow h, H; h(x)}$$

$$\textsc{assign}\quad \frac{H; e \Downarrow h, H'; v}{H; \mathbf{var}\ x = e \Downarrow (x \mapsto v, h), H'; v}$$

To evaluate a quotation, the semantics "switches" from the main big-step judgment to a quoted judgment written $H; h; e \Downarrow_i H'; h'; e'$ where i is the current quotation level. The latter judgment scans over the quoted expression to find escapes that need to be evaluated eagerly with the main semantics. It threads through a heap H, which may be updated inside escaped expressions, and a materialization environment h, which holds the results of top-level materialization escapes. The rules leave most expressions, such as assignments, intact:

$$\textsc{quoted-assign}\quad \frac{H; h; e \Downarrow_i H'; h'; e'}{H; h; \mathbf{var}\ x = e \Downarrow_i H'; h'; \mathbf{var}\ x = e'}$$

The judgment switches back to ordinary $\Downarrow$ interpretation when an escape goes beyond the current quotation level. For example, a splice ${}_n[e]$ where $n = i$ returns to the top level; it evaluates e to a code value and includes the resulting expression:

$$\textsc{quoted-splice-resume}\quad \frac{H; e \Downarrow H'; \langle\!\langle e_q, h_q \rangle\!\rangle \qquad n = i \qquad \text{merge}(h, h_q) = h'}{H; h; {}_n[e] \Downarrow_i H'; h'; e_q}$$

A merge(h, h') helper judgment combines the variable mappings from two environments.

Safety theorem. We state a type safety theorem for BraidCore, which says that well-typed programs do not go wrong.

Theorem 1. (*Type safety*)
If $\cdot \vdash e : \tau; \Gamma$, then $\cdot; e \Downarrow H; v$.

5. Static Staging for Real-Time Graphics

This section describes the design and implementation of BraidGL, a prototype language and compiler for real-time 3D rendering on CPU–GPU systems. The compiler emits a combination of host code in JavaScript, shader code in GLSL [70], and interface code using the WebGL API [39].

5.1. Targets and Annotations

In Braid, programmers choose a *target* for each quote that controls how it is compiled. Annotations on quotes, written t`<`e`>`, select a target t. Targets specify the code-generation language and hardware. Targets are defined in compiler extensions with three components: a language variant that adds platform-specific intrinsics and removes unsupported features; a code-generation backend; and communication strategies that implement materialization from other targets.

The BraidGL compiler extension defines a shader target, written `glsl<`e`>`. The compiler emits shader quotes as GLSL source code in string literals. The `!` operator cannot run shaders directly; instead, BraidGL defines special intrinsic operations that execute the code in the graphics pipeline.

BraidGL also includes a `js<`e`>` annotation, which directs the compiler to emit JavaScript code as a function declaration and to execute it on the CPU. The per-frame render stage uses this code type in place of plain `<`e`>` to avoid the cost of using JavaScript's `eval` on each frame.

5.2. Binding Stages with Intrinsics

BraidGL provides three intrinsics that execute a graphics program's stages. The `render` intrinsic registers code to run on the CPU to draw each frame; `vertex` binds a vertex shader; and `fragment` associates a fragment shader with a vertex shader. We leave less common stages, such as tessellation, geometry, and compute shaders [70], for future work.

The `render` intrinsic registers code that draws each new frame. The `vertex` intrinsic compiles to a call to WebGL's `gl.useProgram()`, which instructs the `gl` context to use a given compiled shader program for the next draw call. Finally, `fragment` instructs the compiler to emit setup code that compiles the fragment shader's code together with its containing vertex shader to create an executable program. Together, this nesting of intrinsics and quotes in BraidGL:

```
render js<
  # (render loop code)
  vertex glsl<
    # (vertex shader code)
    fragment glsl<
      # (fragment shader code) > > >
```

compiles to this JavaScript code:

```
var QUOTE_1 = "(vertex shader code)";
var QUOTE_2 = "(fragment shader code)";
var shader_1 = compile_shader(gl, QUOTE_1, QUOTE_2);
add_renderer(function () {
  // (render loop code)
  gl.useProgram(shader_1); });
```

The `compile_shader` runtime function wraps WebGL's lower-level GLSL compilation calls, and `add_renderer` registers a callback to draw each frame [77]. The end result is a JavaScript program that resembles hand-written WebGL graphics code.

5.3. Binding Shader Parameters

Materialization expressions and cross-stage references in BraidGL denote communication constructs in WebGL and

GLSL. This example communicates a vector value from the CPU to both shader stages:

```
var color = vec3(0.2, 0.5, 0.4);
vertex glsl< ... color ...
  fragment glsl< ... color ... > >
```

The compiler generates three ingredients for each materialization. First, it emits a `uniform`, `varying`, or `attribute` declaration in the GLSL code for each shader that indicates where the value comes from:

```
var QUOTE_1 =
  "uniform vec3 param_1;" +
  "void main() {" +
  "  ... param_1 ... " +
  "}";
```

Next, the compiler emits a location handle lookup:

```
var shader_1 = compile_shader(gl, QUOTE_1, QUOTE_2);
var shader_1_loc_1 =
  gl.getUniformLocation(shader_1, "param_1");
```

Finally, when binding the shader with the `vertex` intrinsic, the compiler emits code to bind the materialized values:

```
gl.uniform3fv(shader_1_loc_1, vec3(0.2, 0.5, 0.4));
```

The materialization's type dictates which OpenGL API call to use. For example, `gl.uniformMatrix4fv` communicates a 4×4 matrix value.

Vertex attributes. The above example shows a *uniform* shader parameter, which is constant across all vertex shader invocations. BraidGL also supports *vertex attribute* parameters for per-vertex data. For example, programs may communicate a model's mesh coordinates using a vertex attribute.

In BraidGL, a `T Array` represents a dynamically-sized buffer of `T` values. When a shader's materialization expression has a `T Array` type, the compiler binds it as an attribute instead of as a uniform. The result in the shader quote has type `T` and refers to the *current* value of the attribute. For example, in this materialization from our earlier example:

```
var position = getPositions(mesh);
vertex glsl<
  ... %[position] ... >
```

the `position` variable has type `Vec3 Array`, but the materialization inside the vertex shader produces a single `Vec3`. The BraidGL compiler uses WebGL's `gl.vertexAttribPointer` API call to bind attribute buffers.

OpenGL does not support direct communication of attributes from the CPU to the fragment shader—attributes can only communicate to the vertex shader. To implement array references in fragment shaders, BraidGL emits code to copy the attribute's value through the vertex shader into the fragment shader via `varying`-qualified parameters.

6. Evaluation

We evaluate Braid's ergonomics to show how static staging helps graphics programmers explore placement and specialization choices. We develop three case studies and describe how to accomplish a variety of well-known visual effects and optimization techniques.

Program	BraidGL LoC	JS+GLSL LoC
Phong	61	119
head	59	141
couch	144	218

Table 2. The lines of code in each case study program and its equivalent JavaScript and GLSL code.

We implement a compiler for Braid and BraidGL in 8028 lines of TypeScript and a grammar for the PEG.js [51] parser generator. BraidGL emits JavaScript code that uses the WebGL 1.0 API, a Web standard supported by all major browsers [39]. The compiler and the generated code all run in the browser, so we have also developed a live-coding environment for BraidGL. (URL omitted for review.)

The implementation also includes a complete interpreter for Braid. Unlike the compiler, the interpreter implements quotes using runtime ASTs, so it can pretty-print staged code. Pretty-printing is useful for debugging Braid programs, but we use the compiler for all performance experiments.

Table 2 shows the lines of BraidGL code in each program compared to the equivalent JavaScript and GLSL, which can be twice as verbose. By design, the BraidGL compiler emits code that matches handwritten WebGL code, up to cosmetic differences such as variable naming. The generated programs have the same performance as their hand-written equivalents.

To evaluate performance, we measure the time it takes to draw each frame. Each program renders an 8×8×8 grid of objects. A test harness loads the scene in a browser and measures the execution time of the compiled JavaScript using the browser's high-resolution timing API, `Performance.now` [32]. We also measure *draw time*, the total time per frame spent in OpenGL's `glDrawArrays` and `glDrawElements` calls, which invoke the graphics pipeline to draw pixels. Draw time measurements help separate the time spent on CPU and communication from the core GPU rendering time. We collect per-frame latencies over 8 seconds and report the mean and 95th percentile times.

The experiments used an Apple MacBook Pro with a quad-core 2.3 GHz Intel 3615QM CPU and an Nvidia GeForce GT 650M GPU running the Safari 10.0 browser with WebKit 12603.1.1 on macOS 10.12 (build 16A270f).

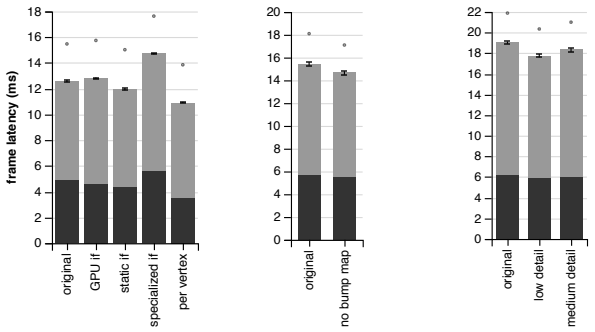
6.1. Phong Lighting Model

The Phong reflection model [61] is a ubiquitous algorithm for approximating lighting effects. Figure 5a shows the Phong program stylizing the "Stanford bunny" mesh [3].

Programmers often render 3D objects with different components of a general shader program. Phong lighting, for example, consists of two components: *diffuse* lighting, which resembles light cast on a matte surface, and a *specular* component, which approximates reflections on a glossy material. Figure 5b shows an object rendered with only diffuse lighting, simulating a matte material. We explore three strategies

(a) Phong lighting. (b) Diffuse only. (c) Per vertex. (d) Head. (e) No bump map. (f) Couch. (g) Medium detail.

Figure 5. Outputs from the case study programs.



(a) Phong lighting. (b) Head. (c) Couch.

Figure 6. The average time to draw a frame. The bar height is the total time, and the black portion is the time spent in WebGL draw calls. Error bars show the standard error of the mean. Points show the 95th percentile latency.

for customizing the Phong shader to draw matte and glossy objects: a straightforward dynamic conditional and two specialization techniques. We then demonstrate a common optimization that trades off accuracy for efficiency.

Dynamic conditional. To choose an object's appearance, the programmer can choose to introduce a dynamic condition into the GPU-side shader code to choose an object's appearance. The CPU chooses whether to draw a given object as matte or glossy, and then the GPU includes one or both lighting components:

```
render js<
  var matte = get_appearance();
  vertex glsl<   # ...
    fragment glsl<   # ...
      var color = if matte diffuse (diffuse + ...);
```

The final `...` placeholder represents the code that computes the secular Phong component. To communicate the `matte` condition variable, the Braid compiler generates code to bind a uniform parameter to the Phong shader. The Braid code is free of boilerplate and makes it clear that the parameter choice is located on the CPU.

Figure 6a shows the dynamic conditional's small performance cost. The *original* bar is the unconditional shader and *GPU if* denotes the version with a dynamic condition. The latter takes 12.8 ms on average to render each frame, compared to 12.6 ms for the original.

Static conditional. Programmers often need to convert dynamic conditions into compile-time decisions for efficiency. When all objects in a scene use the same parameters—when they are all matte, for example—a dynamic conditional is wasteful. A compile-time *static if* construct uses a macro that executes at compile time:

```
var matte = 1;   # Compile-time parameter.
var static_if = fun c:Int t:$<Float3> f:$<Float3> ->
  if c t f;
!<
  render js<
    # ... Phong code ... > >
```

Wrapping the entire program in `!<...>` introduces a compile-time stage. To convert the dynamic condition to a static one, we replace `if` with `@static_if` in the shader code:

```
var color = @static_if matte diffuse (diffuse + ...);
```

No other changes are necessary. The generated WebGL code does *not* use a uniform parameter to communicate `matte`, and the compiler removes the code for specular reflection. In Figure 6a, the *static if* bar shows this version of the Phong program with `matte` turned on. The frame latency is 12.0 ms, which is slightly faster than the original program. Braid's stages provide zero-cost compile-time specialization.

Specialized conditional. Applications can gain performance by generating simpler versions of a shader and choosing between them on the CPU. This transformation reduces the control burden on the GPU, so modern games can generate hundreds of thousands of specialized shader variants [35]. However, rapidly switching between shaders can incur overhead, so the advantage of specializing depends on the hardware and workload. It is important to experiment with both.

Braid makes it simple to switch between the two strategies. In the Phong program, we add a *specializing if* macro, `@spif`, that resembles `@static_if` but is declared in the CPU stage instead of a compile-time stage. We then replace the `if` in the conditional version of the shader with `@spif` to get identical behavior but different performance:

```
var color = @spif matte diffuse (diffuse + ...);
```

The `@spif` macro uses pre-splicing (Section 3.4.1) to produce two condition-free versions of the GLSL shader. The compiled CPU code chooses which to bind based on the `matte` variable. This flexibility highlights how static staging offers a uniform foundation for both *run-time* decisions for placement and *compile-time* decisions for specialization.

In Figure 6a, the *GPU if* and *specialized if* bars compare between a GPU-side dynamic condition and the specialized equivalent. In this case, the overhead of specialization is a disadvantage: its frame latency is 14.7 ms, compared to

12.8 ms for a standard `if`. The difference emphasizes the need for a language that lets programmers rapidly experiment with both versions to make empirical judgments.

Per-vertex promotion. Shader developers use the vertex and fragment shaders to trade off between performance and visual quality. The fragment stage runs many times between each vertex, so while code in the fragment shader is more costly than per-vertex code, it computes more detailed visual effects. The graphics pipeline interpolates results from the vertex stage for the fragment stage. Programmers *promote* fragment computations to the vertex stage to produce cheaper, lower-quality results.

Traditionally, experimenting with promotion requires large code changes. In Braid, materialization makes it trivial. We can promote the Phong lighting computation to the vertex stage by wrapping it in materialization brackets:

```
vertex glsl<  # ...
  fragment glsl<
    var color = %[
      # ... lighting code ... ] > >
```

Figure 5c shows this version's lower-quality output. In Figure 6a, the *per vertex* bar shows the performance advantage: the optimized version's latency is 10.9 ms, compared to 12.6 ms for the original program.

6.2. 3D-Scanned Head Rendering

To show how developers experiment with textures, we use the *head* 3D scanning data from McGuire [55]. It comprises a mesh, a surface texture, and a bump map. We implement a renderer using all three. Figure 5d shows the output. We demonstrate how programmers can use BraidGL to avoid the boilerplate typically associated with textures and to eliminate their cost with specialization.

Texture mapping. Many shaders use GPUs' special support for *texture mapping*, which "paints" a 2D image onto the surface of a 3D object. The object's data includes a mapping between mesh vertices and coordinates in the image. In WebGL, loading images into the GPU's texture memory requires even more boilerplate code than binding ordinary shader parameters. The programmer chooses a numbered *texture unit*, activates it, and binds a texture object before rendering the object:

```
gl.activeTexture(gl.TEXTURE0);
gl.bindTexture(gl.TEXTURE_2D, preloaded_texture);
gl.uniform1i(sampler_loc, 0);
  // unit "0" corresponds to gl.TEXTURE0 above
```

Then, in GLSL, the texture appears as a uniform value to be used with a texture lookup function, `texture2D`:

```
uniform sampler2D sampler;
varying vec2 texcoord;
void main() {
  vec4 color = texture2D(sampler, texcoord); }
```

The BraidGL compiler implements a materialization strategy that eliminates the boilerplate. Materializing a `Texture` value from the CPU stage to a GPU stage automatically generates binding code for a GPU texture unit. In the *head* program, the GPU materializes a CPU-defined variable `tex`:

```
var texcoord = mesh_texture_coordinates(head);
var tex = texture(load_image("head.jpg"));
render js< # ...
  vertex glsl< # ...
    fragment glsl<
      var color = texture2D(tex, texcoord); > > >
```

The compiler chooses texture units for the two textures in *head* and generates the host and shader code to bind them.

Optional bump mapping. Programmers use *bump mapping* to add detail to surfaces by changing the way they reflect light. A bump map texture dictates how much to deflect the normal vector at each point. In the *head* program, the bump map makes the object look more realistic, but it may be unnecessary in some contexts, such as when the object is in the distance. Figure 5e shows the output without bump mapping.

We use a Braid macro to make bump mapping optional in the *head* program. The macro, `@bump?`, takes the bump-map lookup code as an argument and, when a compile-time `use_bump_map` flag is cleared, returns a default zero-deflection vector instead. When bump mapping is disabled, the corresponding texture materialization is eliminated, and the BraidGL compiler does not generate the code that loads binds the bump texture. The frame latency (Figure 6b) improves from 15.4 ms to 14.7 ms with this change.

6.3. Procedural Couch Model

We use a *couch* scene from He et al. [36] that exemplifies graphics applications with many specializable features and complex trade-offs between efficiency and functionality. It uses a repeating leather texture, an ambient occlusion map to simulate shadows, a mask texture that controls darkening due to wear, bump mapping, and specular mapping to modulate reflections. Figure 5f shows the output.

Texture averaging. To reduce detail levels for efficiency, graphics programmers need to selectively disable texture layers. We introduce a CPU-side `average` function that computes the mean color in an image. Shaders can replace a texture lookup call, such as `texture2D(leather, texcoord)`, with code that instead *materializes* the CPU-side `average(leather)` expression. Instead of communicating an entire texture, the specialized program sends a single vector to the GPU. Static staging enables local code changes that affect global interactions between heterogeneous components.

Rapid trade-off exploration. When developing complex effects, graphics programmers need to rapidly explore a space of configurations and check their output visually. BraidGL makes this exploration simple: we can enable and disable contributions to the surface colors in *couch* by specializing away code in the fragment shader stage and replacing texture lookups with `average`. In BraidGL, eliminating code that uses parameters from the CPU automatically removes the corresponding host-side code that sets up and communicates that data. Exploring specialization options yielded

a range of visual effects, including a low-detail version with a flat appearance and a medium-detail version (Figure 5g) that maintains most of the leatherness while preserving some performance benefit. Figure 6c shows the three versions' frame latencies, which range from 17.7 ms to 19.0 ms. Exploring these changes with traditional APIs would require carefully coordinating simultaneous changes to GPU and CPU code.

7. Related Work

Static staging builds on work from multi-stage programming, language abstractions for placement, and GPU languages.

Multi-stage programming. Research on safe staged programming originated with MetaML [72, 74]. Our primary contribution is the distinction between splicing and materialization. Traditional splicing inlines values into generated code, but our materialization performs inter-stage communication *without code generation*. Prior languages conflate these two modes of communication [33, 43].

Braid is most closely related to "heterogeneous" staged languages, where the meta-level language and quoted language differ, such as Terra [23], variants of MetaOCaml [24, 75], pluggable object languages in MetaHaskell [50], and Scala's lightweight modular staging [66]. Quoted domain-specific languages also separate execution contexts: for example, database queries from application code [20, 58]. This prior work prioritizes code generation, so it lacks Braid's materialization concept for runtime communication between stages. Jeannie [37] is more similar: it uses stages to represent glue between Java and C. We extend this basic idea for placement in heterogeneous hardware.

Several type systems address open code in dynamic multi-stage programming [9, 13, 15, 15, 19, 21, 42, 57, 59, 59, 73]. Braid restricts open code so the compiler may generate all code ahead of time. Future work could explore relaxing these restrictions while preserving AOT compilation.

Braid's staging-based macros resemble hygienic macro expansion systems [27, 28, 47, 48]. Like MacroML [30], Braid uses staging to define macros, but unlike MacroML, it defines them using syntactic sugar for function calls.

Placement abstractions. SPMD languages such as X10 [18] and Chapel [17] use place abstractions. Their goal is to expose locality in distributed systems composed of nodes with roughly equal capabilities. The model is insufficient for single-node heterogeneous systems, such as CPU–GPU hybrids, where one unit is subordinate and units have wildly different capabilities and ISAs.

In ML5 [76], a single distributed program expresses computations that execute on many different machines, and a type system with modal logic rules out unsafe sharing between places. Marking each data element with a place is a data-centric dual to Braid's code-centric placement annotations.

We give programmers explicit control over staged execution, whereas some systems *automatically* decompose monolithic programs into phases. In partial evaluation, *binding-time analysis* finds the parts of a program that can be computed eagerly when a subset of the inputs are available. Jørring and Scherlis [40] define staging transformations as compiler optimizations, and the recent $\lambda^{\mathbb{12}}$ hoists computation in a higher-order, explicitly two-stage language [26].

Programming GPUs. The most common general-purpose programming environments for GPUs are CUDA [60] and OpenCL [71]. OpenCL uses a string-based API similar to OpenGL, while CUDA and other *single-source* GPU languages [2, 31, 38, 68] avoid the hazards of stringly-typed interfaces. Like OpenGL, however, these simpler non-graphics languages still lack safe metaprogramming tools for compile-time and run-time specialization. To help address this shortcoming in CUDA and similar languages, we plan to port Braid to a GPGPU backend.

Mainstream GPU shader languages directly reflect the structure of current real-time graphics hardware pipelines: separate kernels are authored for each programmable stage [5, 22, 41, 56, 70]. To communicate, each kernel declares unsafe interfaces that must align across all kernels.

In *rate-based* languages, type qualifiers describe different rates of computation in a single program that spans all stages of the programs' graphics pipeline [29, 36, 63]. Rates play a similar role to stages in Braid: a compiler uses them to split the program into per-stage kernels. Similarly, *import operators* in Spire [36] are analogous to materialization expressions in Braid: they move data between rates. The rate-based and stage-based approaches differ in two important ways. First, while Braid's stages are always explicit, rate qualifiers can be inferred: Spire automatically identifies a value's set of possible rates. Second, Braid uses a single mechanism to describe both placement and specialization, whereas the rate-based languages focus on placement.

Several experimental languages for shading [1, 6, 7, 10, 11, 25, 52–54, 69], like others for GPGPU programming [45, 46], tend to focus on dynamic code generation, whereas our work focuses on ahead-of-time optimization.

8. Conclusion

We hope this paper leaves the reader with three main ideas:

- Heterogeneous programming languages need abstractions for *placement* and *specialization*.
- With extensions for static code generation, *multi-stage programming* can offer a foundation for both concepts.
- Current tools for *real-time graphics* are especially unsafe, verbose, and brittle. Our community has an opportunity to make graphics development less bad.

We see static staging both as a practical solution to the immediate problems with GPU programming and as a sound semantic foundation for emerging heterogeneous systems.

References

[1] LambdaCube 3D. http://lambdacube3d.com.

[2] The OpenACC application programming interface. http://www.openacc.org/sites/default/files/OpenACC_2pt5.pdf.

[3] The Stanford 3D scanning repository. http://graphics.stanford.edu/data/3Dscanrep/.

[4] Jason Ansel, Cy P. Chan, Yee Lok Wong, Marek Olszewski, Qin Zhao, Alan Edelman, and Saman P. Amarasinghe. PetaBricks: a language and compiler for algorithmic choice. In *ACM Conference on Programming Language Design and Implementation (PLDI)*, 2009.

[5] Apple. Metal shading language guide. https://developer.apple.com/library/ios/documentation/Metal/Reference/MetalShadingLanguageGuide/Introduction/Introduction.html.

[6] Chad Austin and Dirk Reiners. Renaissance: A functional shading language. In *ACM SIGGRAPH/EUROGRAPHICS Conference on Graphics Hardware*, 2005.

[7] Baggers. Varjo: Lisp to GLSL language translator. https://github.com/cbaggers/varjo.

[8] Alan Bawden. Quasiquotation in Lisp. In *ACM SIGPLAN Workshop on Partial Evaluation and Program Manipulation (PEPM)*, 1999.

[9] Zine-El-Abidine Benaissa, Eugenio Moggi, Walid Taha, and Tim Sheard. Logical modalities and multi-stage programming. In *Federated Logic Conference (FLoC) Satellite Workshop on Intuitionistic Modal Logics and Applications (IMLA)*, 1999.

[10] Tobias Bexelius. GPipe. http://hackage.haskell.org/package/GPipe.

[11] Kovas Boguta. Gamma. https://github.com/kovasb/gamma.

[12] Kevin J. Brown, Arvind K. Sujeeth, HyoukJoong Lee, Tiark Rompf, Hassan Chafi, Martin Odersky, and Kunle Olukotun. A heterogeneous parallel framework for domain-specific languages. In *International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 2011.

[13] C. Calcagno, E. Moggi, and T. Sheard. Closed types for a safe imperative MetaML. *Journal of Functional Programming*, 13 (3): 545–571, May 2003a.

[14] Cristiano Calcagno, Walid Taha, Liwen Huang, and Xavier Leroy. Implementing multi-stage languages using ASTs, Gensym, and reflection. In *International Conference on Generative Programming and Component Engineering (GPCE)*, 2003b.

[15] Cristiano Calcagno, Eugenio Moggi, and Walid Taha. ML-like inference for classifiers. In *European Symposium on Programming (ESOP)*, 2004.

[16] Hassan Chafi, Arvind K. Sujeeth, Kevin J. Brown, HyoukJoong Lee, Anand R. Atreya, and Kunle Olukotun. A domain-specific approach to heterogeneous parallelism. In *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*, 2011.

[17] Bradford L. Chamberlain, David Callahan, and Hans P. Zima. Parallel programmability and the Chapel language. *International Journal of High Performance Computing Applications*, 21 (3): 291–312, 2007.

[18] Philippe Charles, Christian Grothoff, Vijay Saraswat, Christopher Donawa, Allan Kielstra, Kemal Ebcioglu, Christoph von Praun, and Vivek Sarkar. X10: An object-oriented approach to non-uniform cluster computing. In *ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA)*, 2005.

[19] Chiyan Chen and Hongwei Xi. Meta-programming through typeful code representation. In *ACM SIGPLAN International Conference on Functional Programming (ICFP)*, 2003.

[20] James Cheney, Sam Lindley, and Philip Wadler. A practical theory of language-integrated query. In *ACM SIGPLAN International Conference on Functional Programming (ICFP)*, 2013.

[21] Rowan Davies and Frank Pfenning. A modal analysis of staged computation. In *ACM SIGPLAN–SIGACT Symposium on Principles of Programming Languages (POPL)*, 1996.

[22] Advanced Micro Devices. Mantle programming guide and api reference 1.0. https://www.amd.com/Documents/Mantle-Programming-Guide-and-API-Reference.pdf.

[23] Zachary DeVito, James Hegarty, Alex Aiken, Pat Hanrahan, and Jan Vitek. Terra: A multi-stage language for high-performance computing. In *ACM Conference on Programming Language Design and Implementation (PLDI)*, 2013.

[24] Jason Eckhardt, Roumen Kaiabachev, Emir Pasalic, Kedar Swadi, and Walid Taha. Implicitly heterogeneous multi-stage programming. *New Generation Computing*, 25 (3): 305–336, January 2007.

[25] Conal Elliott. Programming graphics processors functionally. In *Haskell Workshop*, 2004.

[26] Nicolas Feltman, Carlo Angiuli, Umut A. Acar, and Kayvon Fatahalian. Automatically splitting a two-stage lambda calculus. In *European Symposium on Programming (ESOP)*, 2016.

[27] Matthew Flatt. Composable and compilable macros: You want it when? In *ACM SIGPLAN International Conference on Functional Programming (ICFP)*, 2002.

[28] Matthew Flatt. Binding as sets of scopes. In *ACM SIGPLAN–SIGACT Symposium on Principles of Programming Languages (POPL)*, 2016.

[29] Tim Foley and Pat Hanrahan. Spark: Modular, composable shaders for graphics hardware. In *SIGGRAPH*, 2011.

[30] Steven E. Ganz, Amr Sabry, and Walid Taha. Macros as multi-stage computations: Type-safe, generative, binding macros in MacroML. In *ACM SIGPLAN International Conference on Functional Programming (ICFP)*, 2001.

[31] Kate Gregory and Ade Miller. *C++ AMP: Accelerated Massive Parallelism with Microsoft Visual C++*. O'Reilly, 2012. URL http://www.gregcons.com/cppamp/.

[32] Ilya Grigorik, James Simonsen, and Jatinder Mann. High resolution time level 2: W3c working draft. https://www.w3.org/TR/hr-time/.

[33] Yuichiro Hanada and Atsushi Igarashi. On cross-stage persistence in multi-stage programming. In *International Symposium on Functional and Logic Programming (FLOPS)*, 2014.

[34] Johann Hauswald, Yiping Kang, Michael A. Laurenzano, Quan Chen, Cheng Li, Trevor Mudge, Ronald G. Dreslinski, Jason Mars, and Lingjia Tang. DjiNN and Tonic: DNN as a service and its implications for future warehouse scale computers. In *International Symposium on Computer Architecture (ISCA)*, 2015.

[35] Yong He, Tim Foley, Natalya Tatarchuk, and Kayvon Fatahalian. A system for rapid, automatic shader level-of-detail. In *SIGGRAPH Asia*, 2015.

[36] Yong He, Tim Foley, and Kayvon Fatahalian. A system for rapid exploration of shader optimization choices. In *SIGGRAPH*, 2016.

[37] Martin Hirzel and Robert Grimm. Jeannie: Granting Java Native Interface developers their wishes. In *ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA)*, 2007.

[38] Lee Howes and Maria Rovatsou. SYCL specification. `https://www.khronos.org/registry/sycl/`.

[39] Dean Jackson and Jeff Gilbert. WebGL specification. `https://www.khronos.org/registry/webgl/specs/latest/1.0/`.

[40] Ulrik Jørring and William L. Scherlis. Compilers and staging transformations. In *ACM SIGPLAN–SIGACT Symposium on Principles of Programming Languages (POPL)*, 1986.

[41] John Kessenich. An introduction to SPIR-V: A Khronos-defined intermediate language for native representation of graphical shaders and compute kernels. `https://www.khronos.org/registry/spir-v/papers/WhitePaper.pdf`.

[42] Ik-Soon Kim, Kwangkeun Yi, and Cristiano Calcagno. A polymorphic modal type system for lisp-like multi-staged languages. In *ACM SIGPLAN–SIGACT Symposium on Principles of Programming Languages (POPL)*, 2006.

[43] Oleg Kiselyov. MetaOCaml – an OCaml dialect for multi-stage programming. `http://okmij.org/ftp/ML/MetaOCaml.html`.

[44] Oleg Kiselyov. The design and implementation of BER MetaOCaml. In *International Symposium on Functional and Logic Programming (FLOPS)*, 2014.

[45] Andreas Klöckner. Loo.py: Transformation-based code generation for GPUs and CPUs. In *International Workshop on Libraries, Languages, and Compilers for Array Programming (ARRAY)*, 2014.

[46] Andreas Klöckner, Nicolas Pinto, Yunsup Lee, Bryan Catanzaro, Paul Ivanov, and Ahmed Fasih. PyCUDA and PyOpenCL: A scripting-based approach to GPU run-time code generation. *Parallel Computing*, 38 (3): 157–174, March 2012.

[47] Eugene Kohlbecker, Daniel P. Friedman, Matthias Felleisen, and Bruce Duba. Hygienic macro expansion. In *ACM Conference on LISP and Functional Programming*, 1986.

[48] Byeongcheol Lee, Robert Grimm, Martin Hirzel, and Kathryn S. McKinley. Marco: Safe, expressive macros for any language. In *European conference on Object-Oriented Programming (ECOOP)*, 2012.

[49] Chi-Keung Luk, Sunpyo Hong, and Hyesoon Kim. Qilin: Exploiting parallelism on heterogeneous multiprocessors with adaptive mapping. 2009.

[50] Geoffrey Mainland. Explicitly heterogeneous metaprogramming with MetaHaskell. In *ACM SIGPLAN International Conference on Functional Programming (ICFP)*, 2012.

[51] David Majda. PEG.js: Parser generator for JavaScript. `http://pegjs.org`.

[52] Michael McCool, Zheng Qin, and Tiberiu S. Popa. Shader metaprogramming. In *ACM SIGGRAPH/EUROGRAPHICS Conference on Graphics Hardware*, 2002.

[53] Michael McCool, Stefanus Du Toit, Tiberiu Popa, Bryan Chan, and Kevin Moule. Shader algebra. In *SIGGRAPH*, 2004.

[54] Sean McDirmid. Two lightweight DSLs for rich UI programming. `http://research.microsoft.com/pubs/191794/ldsl09.pdf`.

[55] Morgan McGuire. Computer graphics archive, August 2011. `http://graphics.cs.williams.edu/data`.

[56] Microsoft. Direct3D. https://msdn.microsoft.com/en-us/library/windows/desktop/hh309466(v=vs.85).aspx.

[57] Eugenio Moggi, Walid Taha, Zine-El-Abidine Benaissa, and Tim Sheard. An idealized MetaML: Simpler, and more expressive. In *European Symposium on Programming (ESOP)*, 1999.

[58] Shayan Najd, Sam Lindley, Josef Svenningsson, and Philip Wadler. Everything old is new again: Quoted domain-specific languages. In *ACM SIGPLAN Workshop on Partial Evaluation and Program Manipulation (PEPM)*, 2016.

[59] Aleksandar Nanevski and Frank Pfenning. Staged computation with names and necessity. *Journal of Functional Programming (JFP)*, 15: 893–939, November 2005.

[60] John Nickolls, Ian Buck, Michael Garland, and Kevin Skadron. Scalable parallel programming with CUDA. *Queue*, 6 (2): 40–53, March 2008.

[61] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18 (6): 311–317, June 1975.

[62] Phitchaya Mangpo Phothilimthana, Jason Ansel, Jonathan Ragan-Kelley, and Saman Amarasinghe. Portable performance on heterogeneous architectures. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2013.

[63] Kekoa Proudfoot, William R. Mark, Svetoslav Tzvetkov, and Pat Hanrahan. A real-time procedural shading system for programmable graphics hardware. In *SIGGRAPH*, 2001.

[64] Andrew Putnam, Adrian M. Caulfield, Eric S. Chung, Derek Chiou, Kypros Constantinides, John Demme, Hadi Esmaeilzadeh, Jeremy Fowers, Gopi Prashanth, Gopal Jan, Gray Michael, Haselman Scott Hauck, Stephen Heil, Amir Hormati, Joo-Young Kim, Sitaram Lanka, James Larus, Eric Peterson, Simon Pope, Aaron Smith, Jason Thong, Phillip Y. Xiao, and Doug Burger. A reconfigurable fabric for accelerating large-scale datacenter services. In *International Symposium on Computer Architecture (ISCA)*, 2014.

[65] Jonathan Ragan-Kelley, Connelly Barnes, Andrew Adams, Sylvain Paris, Frédo Durand, and Saman Amarasinghe. Halide: A language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines. In *ACM Conference on Programming Language Design and Implementation (PLDI)*, 2013.

[66] Tiark Rompf and Martin Odersky. Lightweight modular staging: A pragmatic approach to runtime code generation and compiled DSLs. In *International Conference on Generative Programming and Component Engineering (GPCE)*, 2010.

[67] Tiark Rompf, Arvind K. Sujeeth, Kevin J. Brown, HyoukJoong Lee, Hassan Chafi, and Kunle Olukotun. Surgical precision JIT compilers. In *ACM Conference on Programming Language Design and Implementation (PLDI)*, 2014.

[68] Ben Sander, Greg Stoner, Siu-Chi Chan, Wen-Heng Chung, and Robin Maffeo. HCC: A C++ compiler for heterogeneous computing. `http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2015/p0069r0.pdf`.

[69] Carlos Scheidegger. Lux: the DSEL for WebGL graphics. `http://cscheid.github.io/lux/`.

[70] Mark Segal and Kurt Akeley. The OpenGL 4.5 graphics system: A specification. `https://www.opengl.org/registry/doc/glspec45.core.pdf`.

[71] John E. Stone, David Gohara, and Guochun Shi. OpenCL: A parallel programming standard for heterogeneous computing systems. *IEEE Design & Test*, 12 (3): 66–73, May 2010.

[72] Walid Taha. *Domain-Specific Program Generation: International Seminar, Dagstuhl Castle, Germany, March 23–28, 2003. Revised Papers*, chapter A Gentle Introduction to Multi-stage Programming, pages 30–50.

[73] Walid Taha and Michael Florentin Nielsen. Environment classifiers. In *ACM SIGPLAN–SIGACT Symposium on Principles of Programming Languages (POPL)*, 2003.

[74] Walid Taha and Tim Sheard. Multi-stage programming with explicit annotations. In *ACM SIGPLAN Workshop on Partial Evaluation and Program Manipulation (PEPM)*, 1997.

[75] Naoki Takashima, Hiroki Sakamoto, and Yukiyoshi Kameyama. Generate and offshore: Type-safe and modular code generation for low-level optimization. In *Workshop on Functional High-Performance Computing (FHPC)*, 2015.

[76] Tom Murphy VII, Karl Crary, and Robert Harper. Type-safe distributed programming with ML5. In *Conference on Trustworthy Global Computing (TGC)*, 2007.

[77] Web Hypertext Application Technology Working Group. HTML living standard. Section 8.9: Animation Frames. `https://html.spec.whatwg.org/multipage/webappapis.html#animation-frames`.